第十三章　特征选择与变换

- 13.1 引言
- 13.2 特征选择(Feature Selection)
- 13.3 特征变换(Feature Transformation)
- 13.4 小结

13.1 引言

- 模式识别中常常把每个对象量化为一组特征来描述，对特征进行处理是模式识别问题的重要步骤
- 通过直接测量得到的特征称为**原始特征**
  - 比如人体的各种生理指标（描述其健康状况）
  - 数字图象中的每点灰度值（以描述图像内容）

13.1 引言

- 原始特征数量可能很大，不利于学习。比如1324*768的256级灰度图像：
  - 直接表示需要786,432 bytes。进行训练识别所需空间、时间、计算量都非常大！
  - 特征有很大的冗余。用少量特征就可以很好地近似表示图像。这与压缩的思想类似。
  - 很少的样本分布在如此高维的空间中，显得十分稀疏，容易产生过学习的现象。维数灾难！

13.1 引言

- 如何提取特征与具体问题有很大关系，特征是**对象的表达**，根据知识来考虑。
  - 特征的稳定性
  - 特征的可分性

- 好的特征胜过好的学习算法！

指纹细节特征


Ridge
Ridge bifurcation
Furrow
Ridge termination
Core
Delta

13.1 引言

- 模式识别中处理特征的方法可分为两类：
  - **特征选择(Feature Selection)**：从原始特征中**挑选**出一些最有代表性、可分性能最好的特征来
  - **特征变换(Feature Transformation)**：希望通过**变换**消除原始特征之间的相关或减少冗余，得到新的特征

13.2 特征选择

13.2 特征选择

- 特征选择从统计的观点来看是变量的选择。
- 特征选择不仅是为了降低特征空间的维数。在很多应用中特征本身具有非常明确的意义，比如基因选择。

13.2 特征选择

- 特征选择是从原始特征中**挑选**出分类性能最好的特征子集来
- 每个特征的状态是离散的 — 选与不选
- 从d个特征中选取r个,共有 C_d^r 种组合。若不限定个数，则共2^d种。—NP 问题
- 这是一个典型的组合优化问题

13.2 特征选择

- 搜索策略
  - 分支定界法
  - **顺序前进法**
  - **顺序后退法**
  - 模拟退火法
  - Tabu 搜索法
  - **遗传算法**

13.2 特征选择

- 顺序前进法——不考虑特征相关性，由少到多，不断增加特征
- 顺序后退法——不考虑特征相关性，由多到少，不断减少特征

13.2 特征选择

- 遗传算法——该算法受进化论启迪，根据“物竞天择，适者生存”这一规则演变
- 几个术语：
  - 基因链码：使用遗传算法时要把问题的每个解编码成一个基因链码。比如要从d个特征中挑选r个,就用一个d位的0或1组成的字符串表示一种特征组合。1表示该特征被选中

    每个基因链码代表一个解，称作一个“个体”，其中的每一位看作一个“基因”

13.2 特征选择

- 群体：若干个体的集合，也就是一些解的集合
- 交叉：选择群体中的两个个体，以这两个个体为双亲作基因链码的交叉，从而产生两个新的个体，作为后代。

X_1　1000　1100　　→　　X'_1　1000 1010

X_2　0100　1010　　　　　X'_2　0100 1100

- 变异：对某个体，随机选取其中一位，将其翻转

1000010　→　1001010

- 适应度：对每个解，以给定的优化准则来评价其性能的优劣，作为其适应度

13.2 特征选择

- 遗传算法的基本框架：
  - 1.初始化进化世代数 t=0
  - 2.给出初始化群体 P(t)，令X_g为任一个体
  - 3.对 P(t) 中每个个体估值，并将群体中最优解X'与X_g比较，若优于X_g，则令X_g= X'
  - 4.如果终止条件满足，则算法结束，X_g为最终结果。否则，转步骤5
  - 5.从P(t)选择个体并进行交叉和变异操作，得到新一代个体P(t+1)，令t=t+1，转步骤3。

13.2 特征选择

- 关于遗传算法的说明：
  - 由步骤3保证了最终解是所搜索过的最优解
  - 常用的终止条件是群体的世代数超过一个给定值，或连续数个世代都没有得到更优解
  - 群体的大小和演化代数是值得重视的参数。在一定范围内，这两个参数大些能得到更好的解
  - 对交叉的亲本选择可采用如下规则：个体的性能越好，被选中的可能性也越大

13.2 特征选择

- 特征选择的方法大体可分两大类：
  - Filter方法：不考虑所使用的分类算法。通常给出一个独立于分类器的选择准则来评价所选择的特征子集S，然后在所有可能的特征子集中搜索出“最优”特征子集。
  - Wrapper方法：将特征选择和分类器结合在一起，即特征子集的好坏标准是由分类器决定的，在学习过程中表现优异的的特征子集会被选中。

13.2 特征选择

- Filter方法的选择准则
  - Fisher判别准则
  - 互信息量准则

13.2 特征选择

- Fisher判别准则——可分性度量

$$J_1 = tr\left(S_w^{-1} S_b\right)$$

$$J_2 = \frac{tr\left(S_b\right)}{tr\left(S_w\right)}$$

$$J_3 = \frac{\left|S_b + S_w\right|}{\left|S_w\right|}$$

13.2 特征选择

- 迭代计算

$$\mathbf{S} = \begin{bmatrix} \tilde{\mathbf{S}} & \mathbf{t} \\ \mathbf{t}^T & s \end{bmatrix}$$

$$\mathbf{S}^{-1} = \begin{bmatrix} \tilde{\mathbf{S}}^{-1} + \frac{1}{d}\tilde{\mathbf{S}}^{-1}\mathbf{t}\mathbf{t}^T\tilde{\mathbf{S}}^{-1} & -\frac{1}{d}\tilde{\mathbf{S}}^{-1}\mathbf{t} \\ -\frac{1}{d}\mathbf{t}^T\tilde{\mathbf{S}}^{-1} & \frac{1}{d} \end{bmatrix}$$

$$d = s - \mathbf{t}^T\tilde{\mathbf{S}}^{-1}\mathbf{t}$$

13.2 特征选择

- 根据每个特征在两类的距离和方差来评价它的分类能力。
  - 准则函数为 $F(j)=\left|\dfrac{\mu_1^j-\mu_2^j}{\sigma_1^j+\sigma_2^j}\right|$
  - 其中 $\mu_1^j,\sigma_1^j,\mu_2^j,\sigma_2^j$ 分别是特征x^j在训练样本中第一类和第二类的均值和标准差。

13.2 特征选择

- 互信息量准则——考虑变量 x^j 和 y 的互信息量。

$$I(j)=\int_{x^j}\int_{y} p(x^j,y)\log\frac{p(x^j,y)}{p(x^j)p(y)}dx^j dy$$

$p(x^j), p(y)$是x^j和y的密度函数，

$p(x^j,y)$是x^j和y的联合密度函数。

对于离散情形，有

$$I(j)=\sum_{x^j}\sum_{y} P(X=x^j,Y=y)\log\frac{P(X=x^j,Y=y)}{P(X=x^j)P(Y=y)}$$

对于连续情形，则需要估计密度函数。

13.2 特征选择

- Wrapper方法
  - 基于最近邻的特征选择
  - 基于SVM的特征选择
  - 基于Fisher判别的特征选择
  - 基于AdaBoost的特征选择

13.2 特征选择

- 基于最近邻的特征选择——OBLIVION
  - 用**顺序后退法**搜索特征子集：从全体特征开始，每次剔除一个特征，使得所保留的特征集合有最大的分类识别率（基于最近邻法)。依次迭代，直至识别率开始下降为止。
  - 用leave-one-out 方法估计平均识别率：用n-1个样本判断余下一个的类别，n次取平均。

13.2 特征选择

- 基于SVM的特征选择——SVM-RFE（ Recursive Feature Elimination)
  - 根据训练得到的SVM线性分类器的系数来判断每个特征的重要性和分类能力。假设由线性SVM得到的分类器为$f(x)=\mathbf{w}^T\mathbf{x}+b=\sum_{i=1}^{d} w_i x^i + b$ 。从全体特征开始，每次剔除一个特征，使得所保留的特征集合有最大的分类识别率。
    - 当w_i 较大时，第i个特征对分类器影响较大；
    - 当w_i 较小时，第i个特征对分类器影响较小；
    - 当w_i 为0时， 第i个特征对分类器几乎没有影响。

13.2 特征选择

- 基于Fisher判别的特征选择——FOM
  - Fisher判别准则

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

$$\mathbf{S}_w \mathbf{w} = \mathbf{m}$$

  - 但是当特征数远远大于样本数时，上面的式子有无穷多个解，我们通过正则化来求解

$$F_1(\mathbf{w}) = \left\| \mathbf{S}_w \mathbf{w} - \mathbf{m} \right\|^2 + \lambda \left\| \mathbf{w} \right\|^2$$

13.2 特征选择

- 我们的目的是进行特征选择，即希望得到的 **w** 最好是由少数非零元素组成。通过引入 $\sigma(\mathbf{w})=\sum_{k=1}^{d}1_{[w_k^2>0]}$， 求解 w 使得下式最小：

$$F_2(\mathbf{w})=\|\mathbf{S}_w\mathbf{w}-\mathbf{m}\|^2+\lambda_1\|\mathbf{w}\|^2+\lambda_2\sigma(\mathbf{w})$$

$\sigma(\mathbf{w})$无法直接求导，我们用$\sum_{i=1}^{d}(1-e^{-\alpha w_i^2})$来逼近，有

$$F(\mathbf{w})=\|\mathbf{S}_w\mathbf{w}-\mathbf{m}\|^2+\lambda_1\|\mathbf{w}\|^2+\lambda_2\sum_{i=1}^{d}(1-e^{-\alpha w_i^2})$$

13.2 特征选择

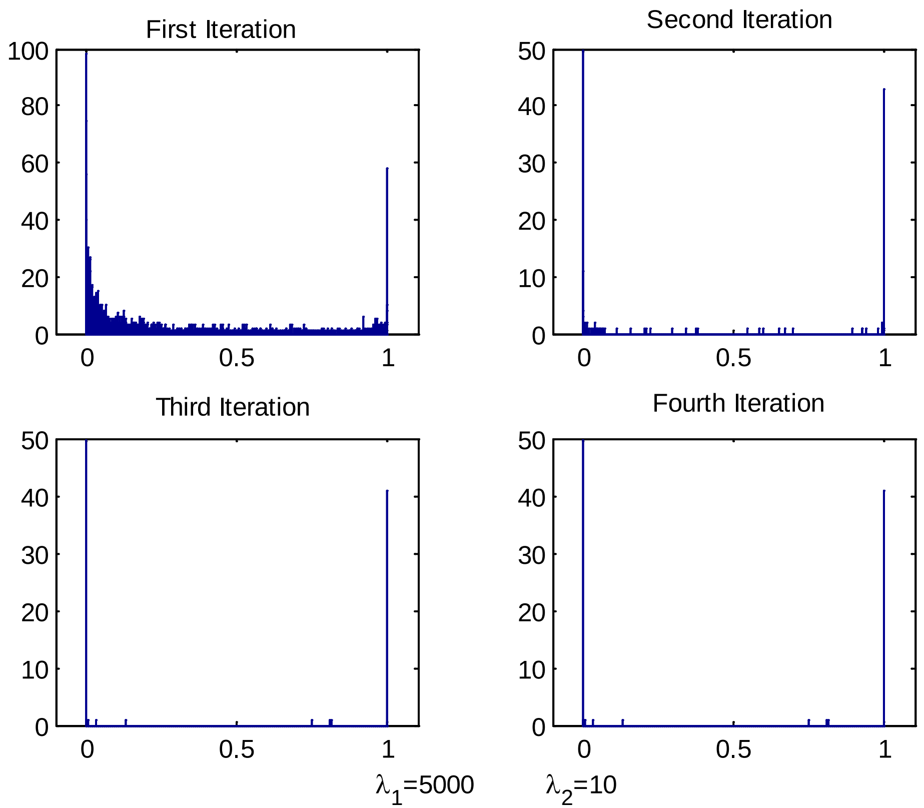
- 迭代求解

$$(\mathbf{S}_w^T\mathbf{S}_w+\lambda_1\mathbf{I}+\lambda_2\alpha\mathbf{D}_i)\mathbf{w}^{i+1}=\mathbf{S}_w^T(\mathbf{m}_1-\mathbf{m}_2)$$

$$\mathbf{D}_i=\begin{pmatrix} e^{-\alpha(w_1^i)^2} & & & \\ & e^{-\alpha(w_2^i)^2} & & \\ & & \ddots & \\ & & & e^{-\alpha(w_d^i)^2} \end{pmatrix}$$

$$w_i=0 \quad if\ 1-e^{-\alpha w_i^2}<T$$

13.2 特征选择


First Iteration
Second Iteration
Third Iteration
Fourth Iteration
100
80
60
50
40
30
20
10
0
0.5
1
λ1=5000
λ2=10

13.2 特征选择

- 基于AdaBoost的特征选择——AdaBoost本质上是从给定有限分类器集合和训练样本集 $H=\{\tilde{h}_j \mid j=1,\cdots,d\}, S$ ，选择合适的分类器进行线性组合。如果我们为每一个特征设计一个分类器，这样分类器选择的过程就实现了特征选择，并且得到最后的分类器。

13.2 特征选择

- 基于AdaBoost的特征选择
  - 首先初始化样本权重
  - 设计每个特征的分类器，如

$$h_j(x^j) = \begin{cases} 1 & \text{if } p_j x^j > p_j \theta_j,\ p_j = \pm 1 \\ -1 & \text{otherwise} \end{cases}$$

  - 根据加权训练样本最小错误率准则选择分类器，也就是选择了特征
  - 调整样本权重
  - 通过循环，最后得到分类器的线性组合

13.3 特征变换

13.3 特征变换

- 特征变换从信号处理的观点来看，是在变换域中进行处理并提取信号的性质，通常具有明确的物理意义。
  - 傅立叶变换
  - 小波变换
  - Gabor变换

13.3 特征变换

- 特征变换从统计的观点来看，就是减少变量之间的相关性，用少数新的变量来尽可能反映样本的信息。
  - 主成分分析PCA （ Principle Component Analysis ）
  - 因子分析FA（Factor Analysis）
  - 独立成分分析ICA （ Independent Component Analysis ）

13.3 特征变换

- 特征变换从几何的观点来看，通过变换到新的表达空间，使得数据可分性更好。
  - 线性判别分析LDA
  - 核方法

13.3 特征变换

- 主成分分析PCA—— $\mathbf{x}=(x^1,x^2,\cdots,x^d)^T$ 是d维随机向量，均值向量和协方差矩阵为

$$\boldsymbol{\mu}=E(\mathbf{x})=\left(E(x^1),E(x^2)\cdots,E(x^d)\right)^T$$

$$\boldsymbol{\Sigma}_{d\times d}=V(\mathbf{x})=E\left[(\mathbf{x}-E(\mathbf{x}))(\mathbf{x}-E(\mathbf{x}))^T\right]$$

$$=\begin{pmatrix} V(x^1) & \mathrm{cov}(x^1,x^2) & \cdots & \mathrm{cov}(x^1,x^d) \\ \mathrm{cov}(x^2,x^1) & V(x^2) & \cdots & \mathrm{cov}(x^2,x^d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(x^d,x^1) & \mathrm{cov}(x^d,x^2) & \cdots & V(x^d) \end{pmatrix}$$

13.3 特征变换

- 样本 $\mathbf{x}=(x^1,x^2,\cdots,x^d)^T$ 可以认为是由观测到的d个变量来描述的。我们希望减少变量之间的相关性，并用少数新的变量来反映样本的信息。

13.3 特征变换

- 随机向量x的协方差矩阵∑的对角元素分别表示x中各分量 $x^1,\cdots,x^d$ 的方差，x的总方差可以为$tr(\Sigma)$ 。

13.3 特征变换

- 我们现在要求线性函数使得新的变量$\mathbf{a}^T\mathbf{x}$的方差尽可能的大，也就是：

$$\max_{\mathbf{a}\neq 0} J(\mathbf{a}) = \max_{\mathbf{a}\neq 0} \frac{V(\mathbf{a}^T\mathbf{x})}{\mathbf{a}^T\mathbf{a}} = \max_{\mathbf{a}\neq 0} \frac{\mathbf{a}^T V(\mathbf{x})\mathbf{a}}{\mathbf{a}^T\mathbf{a}} = \max_{\mathbf{a}\neq 0} \frac{\mathbf{a}^T \Sigma\mathbf{a}}{\mathbf{a}^T\mathbf{a}}$$

$$E(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T E(\mathbf{x})$$

$$\begin{aligned}
V(\mathbf{a}^T\mathbf{x}) &= E\left(\mathbf{a}^T\mathbf{x} - E(\mathbf{a}^T\mathbf{x})\right)^2 = E\left(\mathbf{a}^T\mathbf{x} - \mathbf{a}^T E(\mathbf{x})\right)^2 \\
&= E\left[\left(\mathbf{a}^T\mathbf{x} - \mathbf{a}^T E(\mathbf{x})\right)\left(\mathbf{x}^T\mathbf{a} - E^T(\mathbf{x})\mathbf{a}\right)\right] \\
&= E\left[\mathbf{a}^T\left(\mathbf{x} - E(\mathbf{x})\right)\left(\mathbf{x} - E(\mathbf{x})\right)^T\mathbf{a}\right] \\
&= \mathbf{a}^T V(\mathbf{x})\mathbf{a}
\end{aligned}$$

13.3 特征变换

- 等价于：

$$\max_{\mathbf{a}} J(\mathbf{a}) = \max_{\mathbf{a}} \mathbf{a}^T \Sigma \mathbf{a}$$

$$\mathbf{a}^T \mathbf{a} = 1$$

13.3 特征变换

- 由Lagrange乘子法：

$$\boxed{\begin{aligned} &L(\mathbf{a},\alpha)=\mathbf{a}^T\Sigma\mathbf{a}-\alpha(\mathbf{a}^T\mathbf{a}-1) \\ &\frac{\partial L(\mathbf{a},\alpha)}{\partial\mathbf{a}}=2\Sigma\mathbf{a}-2\alpha\mathbf{a}=0 \end{aligned}}$$

$$\Sigma\mathbf{a}=\alpha\mathbf{a}$$

$$J(\mathbf{a})=\mathbf{a}^T\Sigma\mathbf{a}=\alpha\mathbf{a}^T\mathbf{a}=\alpha$$

13.3 特征变换

- 假设协方差矩阵Σ有r个非零的特征值

$$\Sigma = (\xi_1, \cdots, \xi_p)\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix}(\xi_1, \cdots, \xi_d)^T = \sum_{i=1}^{r} \lambda_i \xi_i \xi_i^T$$

$$\lambda_1 \geq \cdots \geq \lambda_r > 0, \lambda_{r+1} = \cdots = \lambda_d = 0$$

ξ_i是λ_i对应的单位特征向量。$\xi_i^T \xi_j = \delta_{ij}$

$$tr(\Sigma) = \lambda_1 + \cdots + \lambda_r$$

13.3 特征变换

- 协方差矩阵Σ的最大特征值所对应的单位特征向量：

$$\mathbf{a}_1 = \boldsymbol{\xi}_1$$

$$J(\mathbf{a}_1) = \lambda_1$$

13.3 特征变换

- 进一步考虑

$$\max_{\mathbf{a}} J(\mathbf{a}) = \max_{\mathbf{a}} \mathbf{a}^T \Sigma \mathbf{a}$$

$$\mathbf{a}^T \mathbf{a} = 1$$

$$\mathbf{a}^T \mathbf{a}_1 = 0$$

13.3 特征变换

- 由Lagrange乘子法：

$$L(\mathbf{a},\alpha,\beta)=\mathbf{a}^T\Sigma\mathbf{a}-\alpha(\mathbf{a}^T\mathbf{a}-1)-\beta\mathbf{a}^T\mathbf{a}_1$$

$$\frac{\partial L(\mathbf{a},\alpha,\beta)}{\partial\mathbf{a}}=2\Sigma\mathbf{a}-2\alpha\mathbf{a}-\beta\mathbf{a}_1=0$$

$$\beta=\beta\mathbf{a}_1^T\mathbf{a}_1=2\mathbf{a}_1^T\Sigma\mathbf{a}-2\alpha\mathbf{a}_1^T\mathbf{a}$$

$$=2\lambda_1\mathbf{a}_1^T\mathbf{a}-2\alpha\mathbf{a}_1^T\mathbf{a}=0$$

$$\Rightarrow\Sigma\mathbf{a}=\alpha\mathbf{a}$$

$$J(\mathbf{a})=\mathbf{a}^T\Sigma\mathbf{a}=\alpha\mathbf{a}^T\mathbf{a}=\alpha$$

13.3 特征变换

- 协方差矩阵Σ的第二大特征值所对应的单位特征向量：

$$\mathbf{a}_2 = \boldsymbol{\xi}_2$$

$$J(\mathbf{a}_2) = \lambda_2$$

13.3 特征变换

- 因此，我们令 $\mathbf{a}_i = \xi_i, i = 1,2,\cdots,r$ ， 则有

$$\max_{\mathbf{a}^T\mathbf{a}=1} \mathbf{a}^T\Sigma\mathbf{a} = \lambda_1 = \mathbf{a}_1^T\Sigma\mathbf{a}_1$$

$$\max_{\substack{\mathbf{a}^T\mathbf{a}=1 \\ \mathbf{a}^T\mathbf{a}_1=0}} \mathbf{a}^T\Sigma\mathbf{a} = \lambda_2 = \mathbf{a}_2^T\Sigma\mathbf{a}_2$$

$$\vdots$$

$$\max_{\substack{\mathbf{a}^T\mathbf{a}=1 \\ \mathbf{a}^T\mathbf{a}_i=0 \\ i=1,\cdots,r-1}} \mathbf{a}^T\Sigma\mathbf{a} = \lambda_r = \mathbf{a}_r^T\Sigma\mathbf{a}_r$$

13.3 特征变换

- 我们把 $\mathbf{a}_1^T\mathbf{x}, \mathbf{a}_2^T\mathbf{x}, \cdots, \mathbf{a}_r^T\mathbf{x}$ 分别称为随机向量x的第一主成分、第二主成分…第r主成分。它们是r个互不相关的随机变量。它们组成新的随机向量z。

$$\mathbf{W}_{p\times r} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$$

$$\mathbf{z} = (z_1, \cdots, z_r)^T$$

$$= (\mathbf{a}_1^T\mathbf{x}, \cdots, \mathbf{a}_r^T\mathbf{x})^T = \mathbf{W}^T\mathbf{x}$$

13.3 特征变换

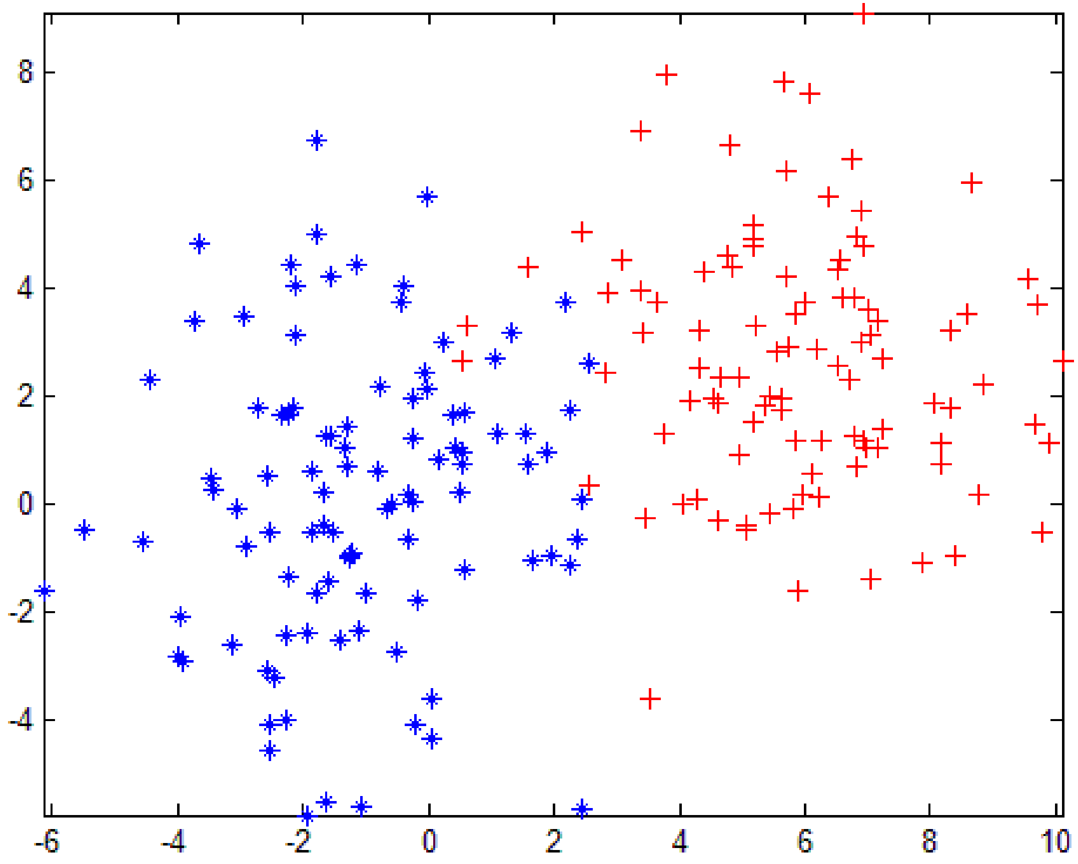
- 总的方差不变

$$V(\mathbf{z}) = \mathbf{W}^T V(\mathbf{x}) \mathbf{W} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{pmatrix}$$

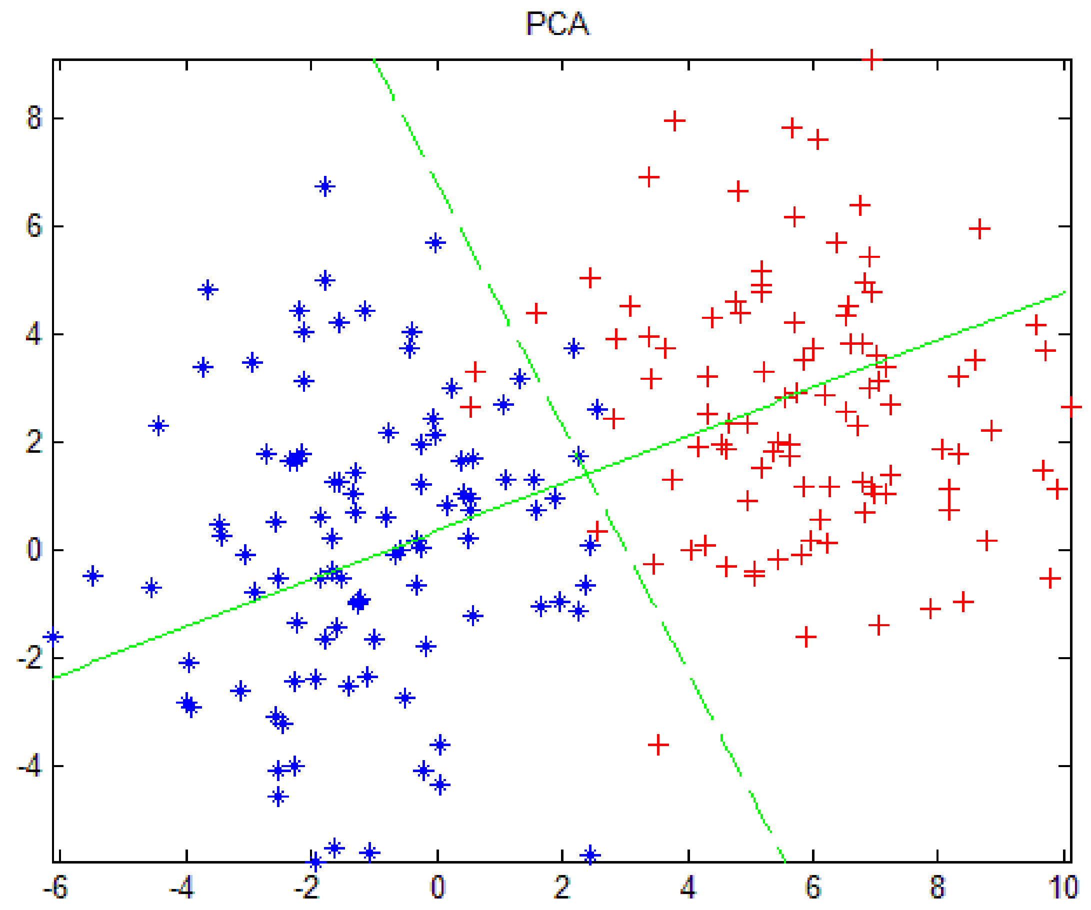
$z_1, \cdots, z_r$互不相关，各自的方差为$\lambda_1, \cdots, \lambda_r$

$\mathbf{z}$的总方差为$\lambda_1 + \cdots + \lambda_r = tr(\Sigma)$

13.3 特征变换



13.3 特征变换



13.3 特征变换

- 因此，主成分分析实际上是把d个随机变量的总方差分解为r个互不相关的随机变量的方差之和，并使第一主成分的方差达到最大，其余的依次递减。

13.3 特征变换

- 实际应用中，我们取前 $p(p \le r)$个特征值对应的特征向量就可以了。

$$\sum_{i=1}^{p} \lambda_i \Big/ \sum_{i=1}^{r} \lambda_i \ge 90\%$$

13.3 特征变换

- 实际应用中，协方差矩阵是未知的，用样本协方差矩阵来估计。

$$\mathbf{X}_{n\times d}=(\mathbf{x}_1,\cdots,\mathbf{x}_n)^T$$

$$\bar{\mathbf{x}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i=\frac{1}{\sqrt{n}}\mathbf{X}^T\mathbf{e},\ \mathbf{e}=\frac{1}{\sqrt{n}}(1,\cdots,1)^T$$

$\mathbf{H}=\mathbf{I}_n-\mathbf{e}\mathbf{e}^T$称为去中心化矩阵。

$$(\mathbf{x}_1-\bar{\mathbf{x}},\cdots,\mathbf{x}_n-\bar{\mathbf{x}})^T=\mathbf{H}\mathbf{X}$$

$$\Sigma_{d\times d}=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T=\frac{1}{n}\mathbf{X}^T\mathbf{H}\mathbf{X}$$

13.3 特征变换

- PCA的计算

1。由去中心化数据矩阵$\mathbf{HX}$的奇异值分解或矩阵$\mathbf{X}^T\mathbf{HX}=n\mathbf{\Sigma}$的特征值$\lambda_i$与单位特征向量$\mathbf{v}_i$（$d$维）分解，取前$r$个特征向量组成矩阵$\mathbf{W}=(\mathbf{v}_1,\cdots,\mathbf{v}_r)$。

2。如果$d>n$，可以由$Gram$矩阵$\mathbf{HXX}^T\mathbf{H}$的特征值l_i与单位特征向量$\mathbf{u}_i$（n维）的分解来计算。

$\mathbf{HXX}^T\mathbf{H}$与$\mathbf{X}^T\mathbf{HX}$有相同的非零特征值，

$$\lambda_i=l_i>0,\ \text{且}\mathbf{v}_i=\frac{\mathbf{X}^T\mathbf{u}_i}{\sqrt{l_i}},\ \mathbf{u}_i=\frac{\mathbf{HXv}_i}{\sqrt{\lambda_i}}。$$

13.3 特征变换

- 线性判别分析LDA （Linear Discriminant Analysis）
  - 在线性判别函数一章，我们讲过Fisher线性判别函数。它的思想是，找一个方向作投影，使得投影后的数据类间距尽可能大，类内距尽可能小。这实际上是两类数据的特征变换，投影到 1 维空间。这一思想可以推广到多类数据，投影到多维空间。

13.3 特征变换

- 假设共有 K 个类 $\{C_1, C_2, \cdots, C_K\}$ ，每类样本数分别为 $\{n_1, n_2, \cdots, n_K\}$ ，定义类内离散度矩阵

样本矩阵 $\mathbf{X}_{n\times d} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T, n = \sum_{j=1}^{K} n_j$

第j类均值向量 $\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$

第j类类内离散度矩阵 $\mathbf{S}_j = \sum_{\mathbf{x}_i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T$

类内总离散度矩阵 $\mathbf{S}_w = \sum_{j=1}^{K} \mathbf{S}_j$

13.3 特征变换

样本均值向量 $\mathbf{m}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i=\frac{1}{n}\sum_{i=1}^{K}n_i\mathbf{m}_i$

总离散度矩阵

$$\mathbf{S}_t=\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T$$

$$=\sum_{j=1}^{K}\sum_{x_i\in C_j}(\mathbf{x}_i-\mathbf{m}_j+\mathbf{m}_j-\mathbf{m})(\mathbf{x}_i-\mathbf{m}_j+\mathbf{m}_j-\mathbf{m})^T$$

$$=\sum_{j=1}^{K}\sum_{x_i\in C_j}(\mathbf{x}_i-\mathbf{m}_j)(\mathbf{x}_i-\mathbf{m}_j)^T+\sum_{j=1}^{K}\sum_{x_i\in C_j}(\mathbf{m}_j-\mathbf{m})(\mathbf{m}_j-\mathbf{m})^T$$

$$=\mathbf{S}_w+\sum_{j=1}^{K}n_j(\mathbf{m}_j-\mathbf{m})(\mathbf{m}_j-\mathbf{m})^T\overset{\Delta}{=}\mathbf{S}_w+\mathbf{S}_b$$

$\mathbf{S}_b=\sum_{j=1}^{K}n_j(\mathbf{m}_j-\mathbf{m})(\mathbf{m}_j-\mathbf{m})^T$ 类间离散度矩阵

13.3 特征变换

- Fisher准则：选择一个最佳投影方向

$$\max_{\mathbf{a}} J_F(\mathbf{a}) = \max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{S}_b \mathbf{a}}{\mathbf{a}^T \mathbf{S}_w \mathbf{a}}$$

13.3 特征变换

- 等价于下面的带等式约束最优化：

$$\max_{\mathbf{a}} \mathbf{a}^T \mathbf{S}_b \mathbf{a}$$

$$s.t. \quad \mathbf{a}^T \mathbf{S}_w \mathbf{a} = 1$$

13.3 特征变换

- 利用Lagrange乘子法：

$$L(\mathbf{a},\alpha) = \mathbf{a}^T \mathbf{S}_b \mathbf{a} - \alpha(\mathbf{a}^T \mathbf{S}_w \mathbf{a} - 1)$$

$$\frac{\partial L(\mathbf{a},\alpha)}{\partial \mathbf{a}} = 2\mathbf{S}_b \mathbf{a} - 2\alpha \mathbf{S}_w \mathbf{a} = 0$$

$$\mathbf{S}_b \mathbf{a} = \alpha \mathbf{S}_w \mathbf{a} \Rightarrow \mathbf{a}^T \mathbf{S}_b \mathbf{a} = \alpha \mathbf{a}^T \mathbf{S}_w \mathbf{a} = \alpha$$

$$\mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{a} = \alpha \mathbf{a}$$

13.3 特征变换

- 假设矩阵 $\mathbf{S}_{\mathbf{w}}^{-1}\mathbf{S}_b$ 有r个非零的特征值

$$\mathbf{S}_w^{-1}\mathbf{S}_b = (\xi_1,\cdots,\xi_d)\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_d\end{pmatrix}(\xi_1,\cdots,\xi_d)^T = \sum_{i=1}^{r}\lambda_i\xi_i\xi_i^T$$

$$\lambda_1 \geq \cdots \geq \lambda_r > 0, \lambda_{r+1} = \cdots = \lambda_d = 0$$

ξ_i是λ_i对应的单位特征向量。$\xi_i{}^T\xi_j = \delta_{ij}$

13.3 特征变换

- 矩阵 $\mathbf{S}_{\mathrm{w}}^{-1}\mathbf{S}_{b}$ 的最大特征值所对应的单位特征向量：

$$\mathbf{a}_1 = \boldsymbol{\xi}_1$$

$$J_F(\mathbf{a}_1) = \lambda_1$$

13.3 特征变换

- 进一步考虑

$$\max_{\mathbf{a}} J_F(\mathbf{a}) = \max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{S}_b \mathbf{a}}{\mathbf{a}^T \mathbf{S}_w \mathbf{a}}$$

$$\mathbf{a}^T \mathbf{a}_1 = 0$$

13.3 特征变换

- 等价于下面的带等式约束最优化：

$$\max_{\mathbf{a}} \mathbf{a}^T \mathbf{S}_b \mathbf{a}$$

$$s.t. \quad \mathbf{a}^T \mathbf{S}_w \mathbf{a} = 1$$

$$\mathbf{a}^T \mathbf{a}_1 = 0$$

13.3 特征变换

- 利用Lagrange乘子法：

$$L(\mathbf{a},\alpha,\beta)=\mathbf{a}^T\mathbf{S}_b\mathbf{a}-\alpha(\mathbf{a}^T\mathbf{S}_w\mathbf{a}-1)-\beta\mathbf{a}^T\mathbf{a}_1$$

$$\frac{\partial L(\mathbf{a},\alpha,\beta)}{\partial\mathbf{a}}=2\mathbf{S}_b\mathbf{a}-2\alpha\mathbf{S}_w\mathbf{a}-\beta\mathbf{a}_1=0$$

$$\mathbf{S}_b\mathbf{a}=\alpha\mathbf{S}_w\mathbf{a}-\beta\mathbf{a}_1\Rightarrow\mathbf{a}^T\mathbf{S}_b\mathbf{a}=\alpha\mathbf{a}^T\mathbf{S}_w\mathbf{a}=\alpha$$

$$\beta\mathbf{a}_1^T\mathbf{S}_w^{-1}\mathbf{a}_1=\alpha\mathbf{a}_1^T\mathbf{S}_w^{-1}\mathbf{S}_w\mathbf{a}-\mathbf{a}_1^T\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{a}$$

$$=\alpha\mathbf{a}_1^T\mathbf{a}-\lambda_1\mathbf{a}_1^T\mathbf{a}=0\Rightarrow\beta=0$$

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{a}=\alpha\mathbf{a}$$

13.3 特征变换

- 矩阵 $\mathbf{S}_{\mathbf{w}}^{-1}\mathbf{S}_b$ 的第二大特征值所对应的单位特征向量：

$$\mathbf{a}_2 = \boldsymbol{\xi}_2$$

$$J_F(\mathbf{a}_2) = \lambda_2$$

13.3 特征变换

- 因此，我们令 $\mathbf{a}_i=\xi_i, i=1,2,\cdots,r$ ，则有

$$\max_{\mathbf{a}^T\mathbf{S}_w\mathbf{a}=1} \mathbf{a}^T\mathbf{S}_b\mathbf{a}=\lambda_1=\mathbf{a}_1^T\mathbf{S}_b\mathbf{a}_1$$

$$\max_{\substack{\mathbf{a}^T\mathbf{S}_w\mathbf{a}=1\\ \mathbf{a}^T\mathbf{a}_1=0}} \mathbf{a}^T\mathbf{S}_b\mathbf{a}=\lambda_2=\mathbf{a}_2^T\mathbf{S}_b\mathbf{a}_2$$

$$\vdots$$

$$\max_{\substack{\mathbf{a}^T\mathbf{S}_w\mathbf{a}=1\\ \mathbf{a}^T\mathbf{a}_i=0\\ i=1,\cdots,r-1}} \mathbf{a}^T\mathbf{S}_b\mathbf{a}=\lambda_r=\mathbf{a}_r^T\mathbf{S}_b\mathbf{a}_r$$

13.3 特征变换

- 因为$\mathbf{S}_b$是K个秩为1或0的矩阵之和，其中只有$K-1$个矩阵是独立的，它的秩最多为$K-1$。因此，对应非零特征根的特征向量最多有$K-1$个,$r \le K-1$。

13.3 特征变换

- 设矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的特征值为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$，假设有$r$个特征值非零，取对应的特征向量作为变换矩阵，则有新的特征向量z。

$$\mathbf{W}_{d\times r} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$$

$$\mathbf{z} = (z_1, \cdots, z_r)^T$$

$$= (\mathbf{a}_1^T\mathbf{x}, \cdots, \mathbf{a}_r^T\mathbf{x})^T = \mathbf{W}^T\mathbf{x}$$

13.3 特征变换

- 我们可以定义关于新的特征向量的类内总离散度矩阵和类间离散度矩阵。

$$\tilde{\mathbf{S}}_w = \sum_{j=1}^{K} \tilde{\mathbf{S}}_j, \tilde{\mathbf{S}}_j = \sum_{\mathbf{z}_i \in C_j} (\mathbf{z}_i - \tilde{\mathbf{m}}_j)(\mathbf{z}_i - \tilde{\mathbf{m}}_j)^T$$

$$\tilde{\mathbf{m}}_j = \frac{1}{n_j} \sum_{\mathbf{z}_i \in C_j} \mathbf{z}_i, \tilde{\mathbf{m}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i = \sum_{j=1}^{k} n_j \tilde{\mathbf{m}}_j$$

$$\tilde{\mathbf{S}}_b = \sum_{j=1}^{K} n_j (\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})^T$$

13.3 特征变换

- 原始特征与新的特征的类内总离散度矩阵和类间离散度矩阵之间的关系：

$$\tilde{\mathbf{m}}_j = \mathbf{W}^T \mathbf{m}_j, \tilde{\mathbf{m}} = \mathbf{W}^T \mathbf{m}$$

$$\tilde{\mathbf{S}}_j = \sum_{\mathbf{z}_i \in C_j} (\mathbf{z}_i - \tilde{\mathbf{m}}_j)(\mathbf{z}_i - \tilde{\mathbf{m}}_j)^T = \mathbf{W}^T \mathbf{S}_j \mathbf{W}$$

$$\tilde{\mathbf{S}}_w = \sum_{j=1}^{K} \tilde{\mathbf{S}}_j = \mathbf{W}^T \mathbf{S}_w \mathbf{W}$$

$$\tilde{\mathbf{S}}_b = \sum_{j=1}^{K} n_j (\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})^T = \mathbf{W}^T \mathbf{S}_b \mathbf{W}$$

13.3 特征变换

- 那么有：

$$\mathbf{W}^T\mathbf{S}_b\mathbf{W} = \left(\mathbf{a}_i^T\mathbf{S}_b\mathbf{a}_j\right)_{r\times r} = \left(\lambda_j\mathbf{a}_i^T\mathbf{S}_w\mathbf{a}_j\right)_{r\times r}$$

$$= \left(\mathbf{a}_i^T\mathbf{S}_w\mathbf{a}_j\right)_{r\times r}\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_r\end{pmatrix} = \mathbf{W}^T\mathbf{S}_w\mathbf{W}\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_r\end{pmatrix}$$

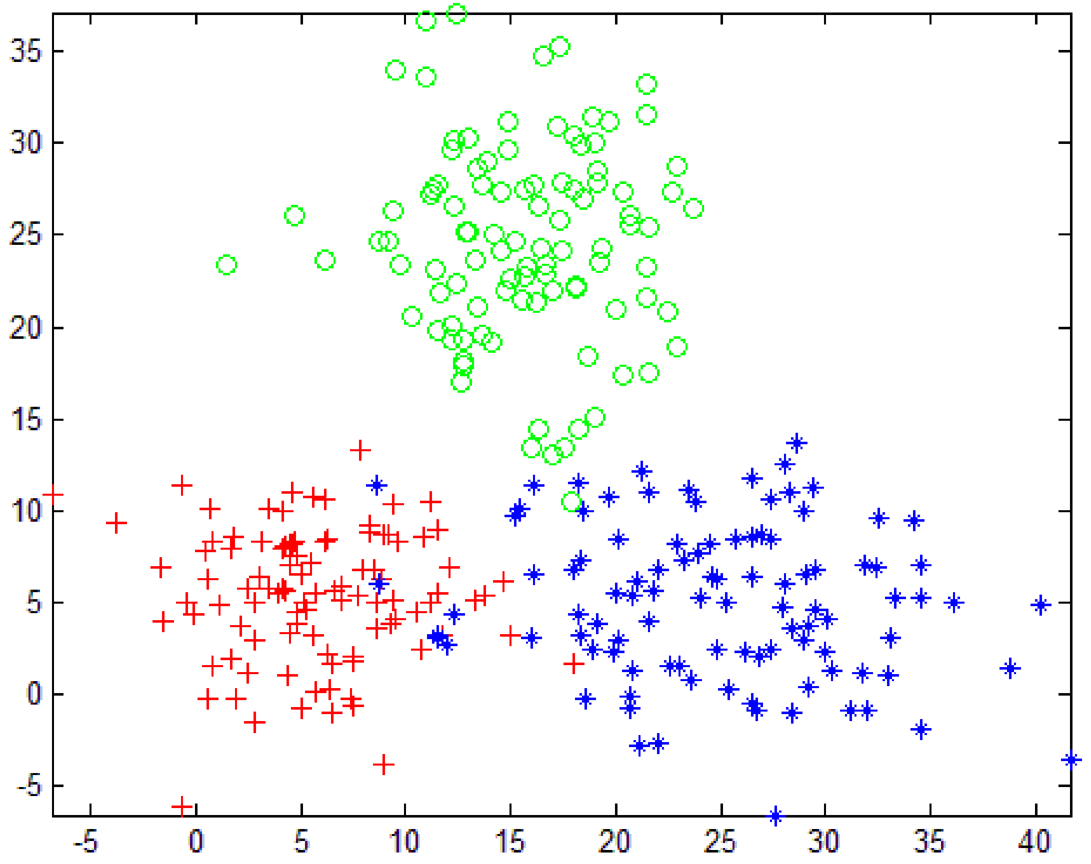
$$\Rightarrow tr\left(\tilde{\mathbf{S}}_w^{-1}\tilde{\mathbf{S}}_b\right) = \max_{\mathbf{W}} tr\left(\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)$$

$$= \sum_{i=1}^{r}\lambda_i = tr\left(\mathbf{S}_w^{-1}\mathbf{S}_b\right)$$

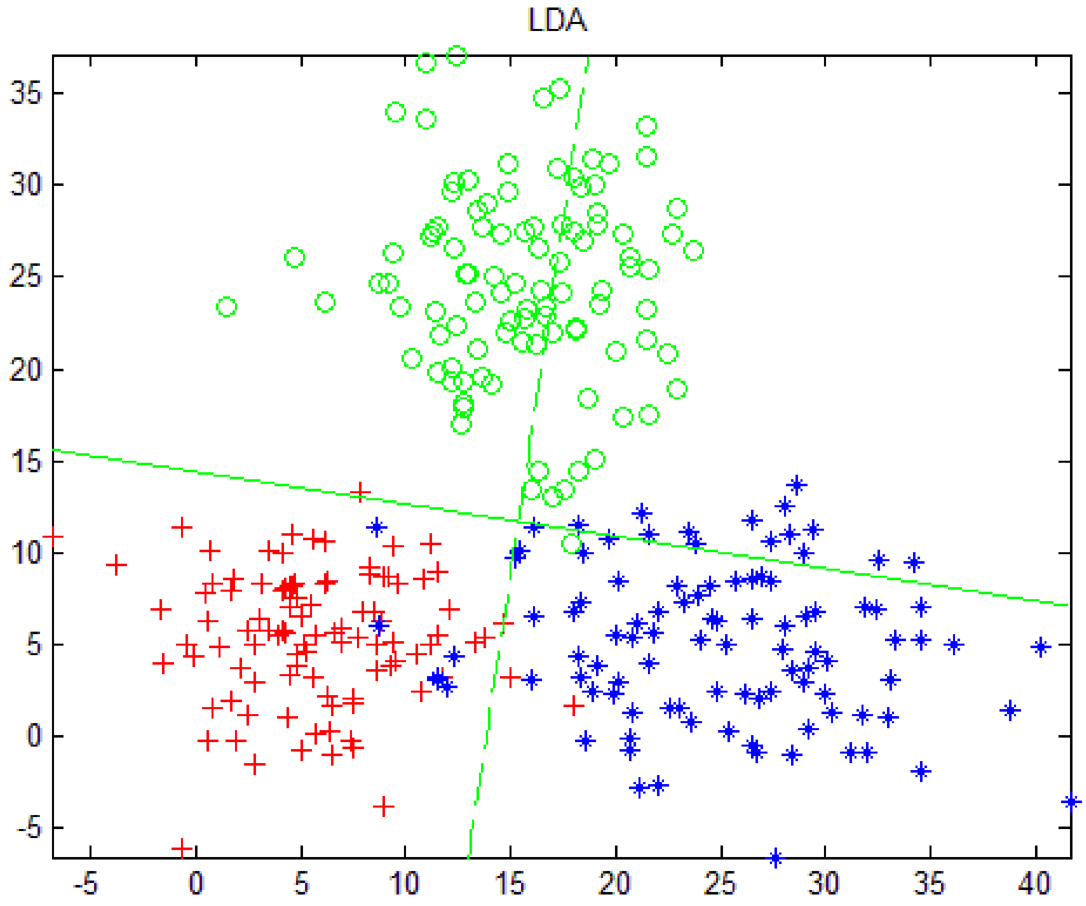
13.3 特征变换

- PCA是通过协方差矩阵的特征值分解得到的特征向量来进行特征变换。没有考虑类别信息。
- LDA引入了类别信息，通过类间离散度矩阵“除以”类内总离散度矩阵$\mathbf{S}_w^{-1}\mathbf{S}_b$的特征值分解得到的特征向量来进行特征变换。

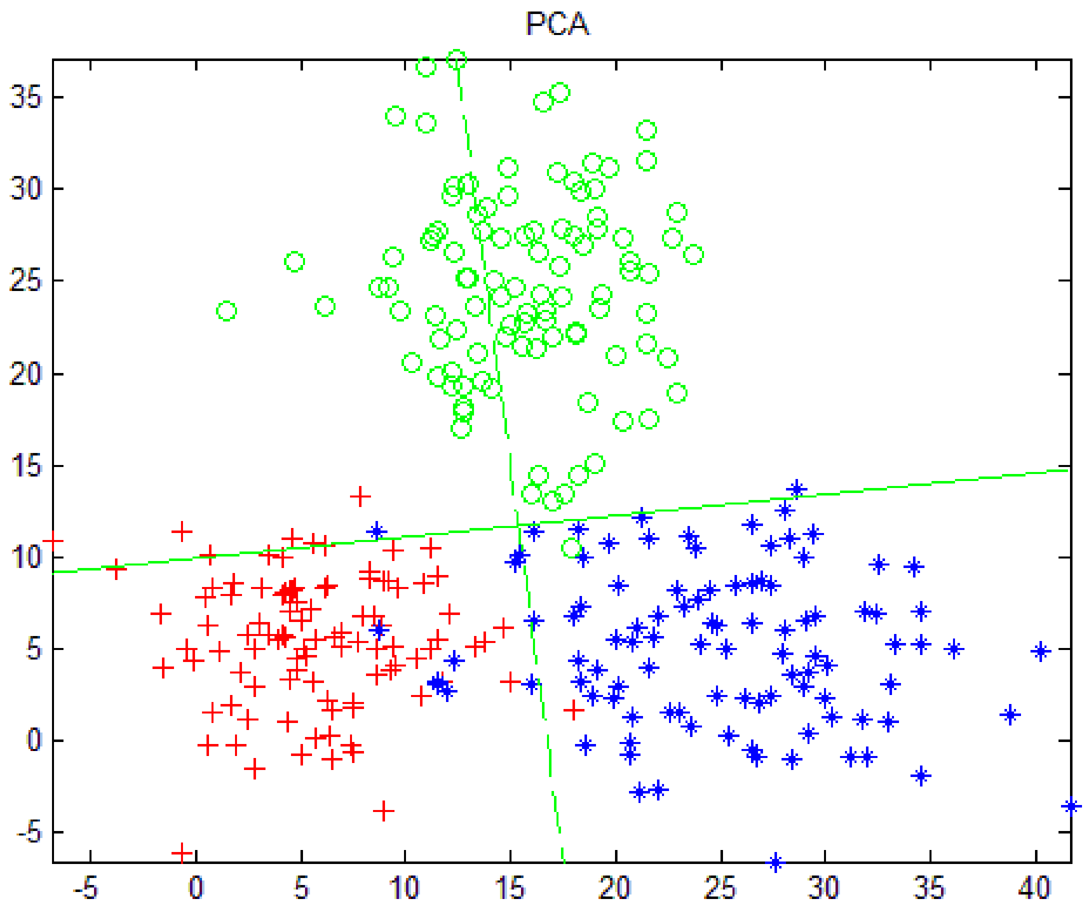
13.3 特征变换



13.3 特征变换



13.3 特征变换



13.4 小结

13.4 小结

- 特征选择是从原始特征中**挑选**出一些最有代表性、可分性能最好的特征来，是一个典型的组合优化问题。不仅可以降低特征空间的维数，特征本身常常具有明确的意义。
  - Filter方法
  - Wrapper方法

13.4 小结

- 特征变换是希望通过**变换**消除原始特征之间的相关或减少冗余，得到新的特征，使得数据可分性更好。
  - PCA
  - LDA

参考文献

- [1] R. Kohavi and G.H. John,Wrappers for feature subsets selection problem, Artificial Intelligence Journal, 97:12, pp.273-324,1997.
- [2] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research, Vol.3, No.7/8, pp.1157-1182, 2003.
- [3] Paul Viola and Michael J. Jones, Robust Real-Time Face Detection, International Journal of Computer Vision, Vol.57, NO.2, pp:137-154,May 2004.
- [4] 张尧庭、方开泰著，《多元统计分析》，科学出版社 1999年。